人工智能在匿名网络追踪中网站指纹的应用综述

王志
中国科学院信息工程研究所，北京
wangzhi@iie.ac.cn

摘要 自从互联网提出以来，网络的匿名性一直是一个公众关注的重要议题。从保证电子选举公平性到保护用户隐私，匿名性起着越来越重要的作用。随着匿名网络的提出，对于公众而言，匿名性得到了一定的保障。但随着匿名对抗技术的不断升级，匿名性也逐渐面临着威胁。本文以匿名性及其保障与对抗技术为主线，综述了匿名网络及其对抗技术的发展。文章主要以网站指纹技术的发展论述人工智能技术在匿名网络中的应用。

关键字 匿名网络，人工智能，Tor，网站指纹

1. 背景

匿名性一直是网络用户的追求之一。生活中有很多对匿名性有很大需求的实际应用场景，如电子选举、机要通信等。随着互联网的发展与普及，互联网用户的隐私意识逐渐增强，用户的隐私问题也受到越来越多的重视，匿名网络随之得到进一步的发展。而自从匿名网络提出以来，针对匿名网络的分析、构造和攻击绵延不绝。研究人员分析并构造许多匿名网络，而攻击者也逐渐对各匿名网络系统进行分析和破解，攻防的博弈战在匿名网络领域拉开。

上世纪80年代已经有一些学者在匿名和隐私方面做了研究。[1]率先从密码学的角度提出了基于节点混淆（Mix）的匿名通信方式，通过混合多个发送者的消息传送给接收者，以达到电子邮件不可追溯的问题。A. Pfitzmann等[2]结合匿名网络的特性，审视匿名网络的结构，给出了匿名性（Anonymity）的分类和定义。从2000年开始，A. Pfitzmann等结合有关隐私问题的研究，在一般的通信系统模型上，给出了通信系统的一系列隐私与匿名的相关术语及定义，并维护至2010年，包括匿名性（Anonymity）、不可链接性（Unlinkability）、不可检测性（Undetectability）、不可观测性（Unobservability）、假名性（Pseudonymity）、可鉴别性（Identifiability）和身份管理（Identity Management）等[3]。

对匿名性进行评估的工作有很多。1999年，Michael K. Reiter等[4]做出了匿名性的量化工作, 他们把匿名性的等级定义为1 – p，其中 p 是攻击者给出的和当前用户相关的概率值。O. Berthold等[5]和C. Diaz等[6]从信息论的角度分别给出了匿名性的量化标准。M. Edman等[7]提出了用二分图的方法来量化匿名性。在定性测量方面，M. Bhargava等[8]和J. Halpern等[9]分别从概率的角度判断通信系统是否匿名，M. Backes等[10]从组合定理的角度评估Tor的路径选择的匿名性。包括匿名性在内，Wang等[11]从用户和开发者的角度，提出了对匿名网络的反追踪性（Anti-traceability）、抗封锁性（Anti-blockade）、抗监听性（Anti-eavesdropping）、健壮性（Robustness）和可用性（Usability）的评估方法。

在保障匿名性的过程中，研究人员提出了一系列的方法。最初阶段，为了防止对流量进行的窃听，提出了对数据进行加密的方法，如使用SSL协议进行通信。但研究人员很快提出了对抗措施。Mistry[12]和Cheng[13]等人率先指出攻击者可以通过对加密流量的分析来确定请求的网站URL。他们认为，对特定的网站，传输的数据流具有明显的特征，并能用以识别URL。不过，这些早期研究基于HTTP/1.0，随着HTTP/1.1的提出[14]，这类攻击已经不再奏效。随后，研究人员提出了更有效的流量分析技术。2001年，Raymond[15]对流量分析进行了综述，介绍了流量分析方法的协议、攻击和设计等。攻击者可以通过监听加密流量获取到关于传输内容和传送方的信息。Bissias等[16]是第一批根据IP数据包的大小、到达时间来进行此类攻击的。他们在Open SSH信道到代理服务器见部署实验设备，计算共计100个网站的数据，得到了检测率为20%的结果。与此同时，Hintz[17]和Sun[18]等人提出了网站指纹（Website Fingerprinting）的方法。他们认为，如果攻击者利用一个事先构造好的“指纹库”，通过对传输数据包大小的分布进行分析，能识别指定的网站。Hintz指出，这种攻击仅局限于小规模的网站，而Sun对10万个网站进行指纹识别，也能达到75%的识别正确率。

此外，随着对匿名需求的增加，研究人员提出了不同于寻常网络的网络结构，如Tor[19]、JAP[20]等多跳通信系统（Multi-hop communication system）。Tor通过类似于洋葱的结构，通过至少三跳（三层加密），保护传输数据的安全性，并通过遍布全球的几千个节点[21]，来实现对访问者身份的保护。上述工作中仅针对单加密信道机进行攻击，未考虑Mix或洋葱路由网络，故以上方法在这里并不适用。不过，随着技术的发展，研究人员也将注意力转移到Tor中，在针对Tor的流量分析的研究上取得了一定的成果。Murdoch等[22]、Abbott等[23]、Bauer等[24]对Mix网络和洋葱网络进行了流量分析，取得了一定的成果。

与此同时，统计学习和人工智能的发展也为流量分析和网站指纹带来了新的研究方向。研究人员把机器学习方法应用到网站指纹的领域中，得到了较好的效果。与之前使用相关系数的工作不同，Liberatore等[25]采用了Jaccard系数和

朴素贝叶斯分类器（Naïve Bayes Classifier，NB分类器）进行流量分析，在OpenSSH 信道中能从加密流量中得到特征模式，并达到了73%的准确率（[16]中仅为20%）。统计学习的方法在此领域的效果促使了研究人员的跟进。

2. 应用

D. Herrmann 等[26]采用了多项式朴素贝叶斯分类器（Multinomial Naïve-Bayes Classifier， MNB 分类器）进行网站指纹攻击。这篇文章可以说是针对Tor等匿名网络进行网站指纹的“始祖”。作者调研识别了包括 Open SSH、Open VPN、Stunnel、 思科 IP-Sec VPN 和 Tor、 JAP 环境下的 775 个网站，并应用了 Jaccard 系数、朴素贝叶斯分类器和 MNB 分类器进行训练和识别。对于 Open SSH、 Open VPN 这类单跳系统，作者的识别率能达到 90% 以上；而对于 Tor 和 JAP 这种多跳网络，识别率较低，分别为 JAP 的19.97%和 Tor 的2.96%。

Jaccard系数是一种集合间相似度度量的方法，常用于非监督学习的任务中，也可用于分类问题。在作者进行研究之前，Jaccard系数已在[25]中对单跳系统的网站指纹取得了较好的效果。NB分类器广泛用于监督学习中，并假设集合的各属性之间相互独立。尽管在实际情况中，此假设往往是不成立的，但NB方法还是在分类任务中取得了较大的成功。作者在这里对训练数据采用高斯核密度估计，在实际应用中较为有效的同时，也带来了巨大的计算量。MNB分类是数据挖掘领域的一个经典方法，在半自动化的分类任务如垃圾邮件识别等方面有着较为成功的应用。NB分类器用高斯核估计类别的概率，并选择可能概率最大的分类，而MNB分类器根据所有数据包大小的分布来进行选择。作者收集了数据包的大小、方向、速度等，并统计了TF-IDF，进行余弦归一化。作者的结论显示，对单跳系统，对数据进行TF转换和余弦归一后的效果最好，能达到94.94%-97.64%不等的准确率；对多跳网络（Tor和JAP），对数据只进行余弦归一后的效果最好，分别能达到2.96%和19.97%的准确率。显然，这里的结论认为，匿名网络能抵抗网站指纹攻击。

A. Panchenko等[27]在[26]的基础上进行了改进。作者认为，面对网站指纹，Tor这类匿名网络并不像想象中那么安全。作者率先将流量的数量、时间、方向等信息综合在一起，运用支持向量机（Support Vector Machine，SVM）进行模型训练，并在[26]的数据集上得到了Tor为55%、JAP为80%的准确率。此外，作者将此成果应用在开放环境（Open-World）的场景中，得到了整体TPR（True Positive Rate）为73%、FPR（False Positive Rate）为0.05%的结果。随后，作者对网站指纹进行了混淆，经过混淆对抗后的识别率分别为Tor的3%和JAP的4%。作者首先提供了在开放环境进行网站指纹的方法，这一点十分重要，因为最终

的结果必将用于实际的场景中。如果对实验场景加以各种假设和限制，那么即使有较高的实验效果和提升，也无法进行应用，只属于“纸上谈兵”。

SVM是一种监督学习方法，在数据分类问题上有着较高的准确率。SVM的核心思想是把数据映射成高维向量，并找到合适的超平面进行数据的分类。运用在网站指纹上时，需要获取页面的特征和原始数据，并表示成一个向量。这里作者使用了径向基函数（Radial Basis Function，RBF）作为核函数，并得到最佳的 C 和 γ 值。作者收集的数据包括数据包长度、时间、方向、序列信息、总字节数、数量等信息，并获取Sexually Explicit、Alexa Top Ranked和Alexa Random数据信息，构造开放环境的数据集。如上所述，因识别率的增高，作者最终得到关于Tor和JAP没有那么安全的结论。作者还进行了混淆对抗实验。作者通过填充虚拟流量、模拟访问等手段进行欺骗，最终得到较低的识别率。尽管这个实验只是概念证明类的测试，但还是为以后针对网站指纹的对抗提供了思路和经验。

T. Wang等[28]继续在此基础上进行改进。作者根据网站指纹过程中使用的统计学习方法的不同，将以往的网站指纹方法分为基于非距离的方法和基于距离的方法两种。[26]和[27]分别属于基于非距离的方法和基于距离的方法。从效果上看，作者比较认同基于距离的方法，并在本文中继续采用 SVM 方法进行分类。与以往工作不同的是，作者比较看重指标的选取。作者借鉴了 Cai 等在[29]中提出的最佳弦对齐距离（Optimal String Alignment Distance，OSAD）来测量两个流量间的距离。OSAD最初用来进行词匹配，关注一个实例通过插入、删除、替换、转换等操作形成另一个实例的操作数量，以此作为二者的距离。DL距离（Damerau-Levenshtein distance,DLD）和OSAD十分相近，主要区别在于取消了对转换的限制。因此，当操作成本相同时，DLD不会大于OSAD。作者关注的其他距离指标还有不同的传入/传出数据包代价、不同的传输代价、描述时间的快速类L距离（Fast Levenshtein-like Distance）等。作者在数据采集上也做了优化，从网站的加载上提取了部分数据，作为特征。作者以[29]中的方法和自己修改后的OSAD、快速类L距离进行封闭环境（Closed-World）和开放环境中的对比实验。从实验效果上看，作者修改后的OSAD方法无论是在[29]中的数据集上还是在本文采集的数据集上，准确率均比[29]的方法和快速类L距离高。其中封闭环境的准确率为91%，开放环境的准确率超过90%，召回率为96.9%。作者认为，在开放环境（Alexa Top 1000）中的网站指纹能达到高于95%的召回率，Tor仍需要在保护用户隐私方面进行改进。

紧接着，Wang等[30]又提出了基于 k 近邻（k-Nearest Neighbor，k-NN）分类器的网站指纹方法，比自己上一个工作的准确率更高。k-NN是一个有监督学习方法。相比于其他方法，k-NN有着适合本工作的优势。首先k-NN的训练离不开

距离的计算，这与上文所述作者的思路和工作相吻合。其次，*k*-NN的训练时间比较短，主要在于计算两点之间的距离。另外，*k*-NN可以准确对多模型集合进行分类，而分类器仅需要了解训练集的一个模式即可。作者在本次实验中采用的特征包括以下几个类别。其一是通用特征，包括传输数据总大小、传输总时间、传入/传出数据包总数量等。其二是标记信息，把数据包大小在数据集内的数据包标记为1，不在的标记为0。其三是数据包序列信息，对传出数据包按序添加一个能表示它之前数据包数目的功能和表示此传出数据包与前一个传入数据包之间的数据包总数的功能。其四是每30个数据包的非重叠跨度中的传出数据包的数量。其五是突发数据包的信息。最后是初始数据包，作者将每个序列各方向的前20个数据包的长度作为特征之一。作者将这项工作应用在实际关注的100个网站中，令客户端进行访问，得到了TPR为85%、FPR为0.6%的结果。此外，作者在这里还进行了基于Tor的防护测试，作了关于敌手了解的信息的假设，根据敌手了解的信息的多少来进行下一步的判定。在业界，本文的主要贡献还是在于提出了基于*k*-NN的网站指纹方法。

不过，与此同时，Juarez等[31]研究人员对网站指纹的有效性产生了质疑。他们认为这些网站指纹方法只适用于实验环境中，并不能在实际应用中起到效果。作者认为这些网站指纹工作有一些明显的限制性条件。在客户端配置方面，这些网站指纹工作主要在封闭环境中进行，研究人员往往只假设用户访问指定的*k*个网站，而*k*在实验中的取值往往又很小，很难和广阔的互联网相匹配。即使有些研究工作[27, 28, 29]声称在开放环境进行了测试，也只要求用户访问指定的k个网站之外的网站。在浏览器行为方面，一些工作给定用户访问行为，例如不停打开不同页面，但只打开一个选项卡，这和实际用户的使用大相径庭，获取到的数据也不符合实际情况。此外，在[29]中，作者因实验需要，搭建了许多网站，但这些网站使用的是同一个网站模板，这令实验的可信度大打折扣。在敌手模型方面，研究人员往往假设敌手能检测出不同网站加载的始终，而这个被证实在实际情况下是十分困难的。研究人员也往往假设敌手能从背景流量中区分出所需流量，而一般计算机设备中使用网络的软件有很多，很难从中识别出特定的流量。关于复现性，研究人员总假设敌手可以在被监视者一样的环境中训练模型，这一点在实验中比较常见，但在实际场景中几乎是不可能的。作者并非口说无凭。针对上述假设，作者使用[27, 28, 29]中的方法进行了实验。作者对Alexa Top 100、Top 1000、Top 10000的网站的主页、其他页面和不在此范围内的混淆网站进行数据抓取，构造了新的数据集。经过实验验证，这些方法均达不到声称的准确率。在综合训练时间等代价后，作者认为，在实际应用中，上述的一些假设仍是网站指纹需要面对的问题。

2016年，Panchenko等[32]提出了一种新的网站指纹攻击方法（简称CUMUL）。CUMUL基于一种“微妙”的方法，将网络轨迹映射成一个类表示。作者通过生成网络轨迹追踪的累积行为表示来抽象网页的加载过程，并由此为分类器提取特征。这样隐含地覆盖了其他分类器需要考虑的流量特征，例如分组排序或突发行为。通过设计，作者的分类器可以抵御带宽、拥塞和网页加载时间的差异。这种方法在分类准确性方面优于已有的分类器，同时效率更高。为了在尽可能接近实际的情况下评估网站指纹攻击的有效性，作者构建了一个本领域较有代表性的数据集，由包括主页和其他页面在内的30万个网站数据构成，是以往工作中数据集的数倍之大。在构建数据集的考虑方面，作者并不局限于Alexa数据，而是综合考虑了各种互联网流量，包括社交网站、新闻网站、BT种子网站、知识网站、娱乐网站、博客网站、非英语语系网站、在线数据库和成人网站等。CUMUL通过添加传出数据包分组的长度并减去传入数据包分组的长度来计算累积和，连同总的传入和传出的数据包和字节数，作为SVM的特征。最终，作者用104个特征来表示一个流量实例。在数据集上的实验结果表明，作者的方法具有90%-93%的成功率。

2018年，Rimmer等[33]结合深度学习的方法，提出了网站指纹新的思路。作者认为传统统计学习方法注重于特征选择，从而对流量的特征变化较为敏感，而且大都只在封闭环境下进行测试，在实际使用中仍有局限性。为此，作者提出了基于深度学习的自动化网站指纹方法，分别采用堆栈去噪自编码器（Stacked Denoising Autoencoder，SDAE）、卷积神经网络（Convolutional Neural Network，CNN）和长短期记忆（Long-Short Term Memory，LSTM）网络，通过学习原始流量进行特征提取构建指纹，进而对网站进行识别。

自编码器（Autoencoder，AE）是一种前馈网络，专门用于通过降维来进行特征学习。将多个AE堆积形成深层模型，可以对输入数据的最显着特征进行分层提取，并根据派生特征进行分类，使得SDAE模型能够用于网站指纹。CNN由一系列卷积层构成。卷积层也用于特征提取，从第一层的低级特征开始，逐级抽象。卷积层学习多种过滤器，这些过滤器可以显示出输入数据中包含特定特征的区域，然后对这些输入实例进行降采样，并保留特殊区域。这样，CNN可以检索最重要的特征来进行分类。SDAE需要逐块进行预训练，而CNN不需要大量的预处理。LSTM分类器是特殊类型的循环神经网络（Recurrent Neural Network，RNN），具有增强记忆能力。LSTM的设计允许学习数据中的长期依赖关系，使分类器能够解释时间序列。此处输入的流量轨迹基本上是Tor单元的时间序列，这些动态时间序列很可能包含网站指纹。作者通过Tor和Tor浏览器收集了360万个对Alexa Top 1200的访问数据，并经过筛选，形成了4个主要的数据集CW100、CW200、CW500和CW900，分别包含了筛选后的100个、200个、

500个和900个网站的各2500次访问数据，作为封闭环境的训练集和测试集。作者对Alexa Top 400, 000的网站进行访问，收集400, 000条数据，对CW200中的网站分别进行了2000次访问，收集400, 000条数据，这800, 000条数据构成开放环境的数据集。测试结果表明，作者的方法能达到94%-96%的准确率。

3. 总结与展望

本文从匿名性的提出和匿名网络的发展开始，综述了在这个过程中人工智能方法在流量分析和网站指纹的应用。从[26]开始，网站指纹攻击的研究便打开了大门，而从各类实验也能看出，我们一向认为安全的Tor等匿名网络也不是那么安全。世界上没有绝对的安全，使用Tor也不能带来完全的匿名。随着时间和技术的发展，我认为针对深度学习的对抗也终将运用于Tor网络中。攻防是场长期的博弈战，此起彼伏，此消彼长。而在攻防的对抗中，技术将始终是核心，并得以不断升华。

参考文献

1. Chaum, D.: Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms, In: Communications of the ACM, Vol. 24, No. 2, pp. 84-88 (1981).
2. Pfitzmann A, Waidner M.: Networks without user observability. Computers & Security, 6(2), 158-166 (1985).
3. Pfitzmann A, Hansen M.: A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management, 34 (2010).
4. Reiter, Michael K, and A. D. Rubin.: Crowds: Anonymity for Web Transactions. In: ACM Transactions on Information & System Security1.1, pp. 66-92 (1998).
5. Berthold, Oliver, Pfitzmann, et al.: The disadvantages of free MIX routes and how to overcome them. In: International Workshop on Designing Privacy Enhancing Technologies Design Issues in Anonymity & Unobservability, 63(s164), pp. 30-45 (2001).
6. Díaz C., Seys S., Claessens J., et al.: Towards Measuring Anonymity. In: International Conference on Privacy Enhancing Technologies, pp. 54-68. Springer-Verlag (2002).
7. Edman M, Sivrikaya F, Yener B.: A Combinatorial Approach to Measuring Anonymity. Intelligence and Security Informatics, IEEE, 356-363 (2007).
8. Bhargava M, Palamidessi C.: Probabilistic Anonymity. CONCUR 2005 – Concurrency Theory. Springer Berlin Heidelberg (2005).

9. Halpern J Y, O'Neill K R.: Anonymity and information hiding in multiagent systems. IOS Press (2005).
10. Backes M, Kate A, Meiser S, et al.: (Nothing else) MATor(s): Monitoring the Anonymity of Tor's Path Selection. In: ACM SIGSAC Conference on Computer & Communications Security, pp. 513-524. ACM (2014).
11. Zhi Wang, Jinli Zhang, Qixu Liu, Xiang Cui.: Practical Metrics for Evaluating Anonymous Networks, In: 1st International Conference on Science of Cyber Security, Springer (2018)
12. Mistry, S., Raman, B.: Quantifying Traffic Analysis of Encrypted Web-Browsing, project paper, University of Berkeley (1998).
13. Cheng H., Avnur R.: Traffic Analysis of SSL Encrypted Web Browsing, (1998).
14. R. Fielding, J. Gettys, et al.: RFC 2616 Hypertext Transfer Protocol – HTTP/1.1, (1999).
15. Raymond, Jean François.: Traffic Analysis: Protocols, Attacks, Design Issues, and Open Problems. In: International workshop on Designing privacy enhancing technologies: design issues in anonymity and unobservability Springer-Verlag New York, Inc. pp. 10-29, (2001).
16. Bissias, G. D., Liberatore, M., Jensen, D., & Levine, B. N.: Privacy vulnerabilities in encrypted HTTP streams. In: International Conference on Privacy Enhancing Technologies, Springer-Verlag, Vol.3856, pp.1-11, (2005).
17. Hintz, A.: Fingerprinting Websites Using Traffic Analysis. In: International Conference on Privacy Enhancing Technologies. Springer-Verlag. Vol.2482, pp.171-178, (2002).
18. Q. Sun, Simon, Daniel, R., Wang, et al.: Statistical Identification of Encrypted Web Browsing Traffic. In: IEEE Symposium on Security and Privacy, pp.19-30, (2002).
19. Dingledine R., Mathewson N., Syverson P.: Tor: the second-generation onion router. Journal of the Franklin Institute, 239(2), pp. 135-139 (2004).
20. JAP – ANONYMITY & PRIVACY, https://anon.inf.tu-dresden.de/index_en.html.
21. Tor Metrics Portal, https://metrics.torproject.org.
22. Murdoch, Steven J., and G. Danezis.: Low-Cost Traffic Analysis of Tor. In: IEEE Symposium on Security & Privacy IEEE, pp.183-195, (2005).
23. Abbott, Timothy G., et al.: Browser-Based Attacks on Tor. In: Privacy Enhancing Technologies, International Symposium, Pet 2007 Ottawa, Canada, Revised Selected Papers DBLP, pp.184-199, (2007).
24. Bauer, Kevin, et al.: Low-resource routing attacks against tor. In: ACM Workshop on Privacy in the Electronic Society, Wpes 2007, October DBLP, pp.11-20, (2007).
25. Liberatore, M., & Levine, B. N.: Inferring the source of encrypted HTTP connections. In: ACM Conference on Computer and Communications Security, pp.255-263, (2006).

26. Herrmann, D., Wendolsky, R., and Federrath, H.: Website Fingerprinting: Attacking Popular Privacy Enhancing Technologies with the Multinomial Naïve-Bayes Classifier. In: CCS 2009, Cloud Computing Security Workshop, pp.31-42, (2009).
27. Panchenko, Andriy, et al.: Website Fingerprinting in Onion Routing Based Anonymization Networks, In: ACM Workshop on Privacy in the Electronic Society ACM, pp. 103-114, (2011).
28. T. Wang, and I. Goldberg.: Improved website fingerprinting on Tor. In: ACM Workshop on Workshop on Privacy in the Electronic Society ACM, pp. 201-212, (2013).
29. Cai X, Zhang X C, Joshi B, et al.: Touching from a Distance: Website Fingerprinting Attacks and Defenses, In: ACM Conference on Computer and Communications Security, ACM, pp. 605-616, (2012).
30. T. Wang, X. Cai, R. Nithyanand, R. Johnson, and I. Goldberg.: Effective Attacks and Provable Defenses for Website Fingerprinting, In: USENIX Security Symposium. USENIX Association, pp. 143–157, (2014).
31. Juarez, Marc, et al.: A Critical Evaluation of Website Fingerprinting Attacks, In: ACM SIGSAC Conference on Computer and Communications Security ACM, pp. 263-274, (2014).
32. A. Panchenko, F. Lanze, A. Zinnen, M. Henze, J. Pennekamp, K.Wehrle, T. Engel.: Website fingerprinting at internet scale, In: Network and Distributed System Security Symposium (NDSS). IEEE Computer Society, pp. 1–15, (2016).
33. Rimmer, Vera, et al.: Automated Website Fingerprinting through Deep Learning. In: Network and Distributed System Security Symposium, (2018).